

ANALYSE DES CORRESPONDANCES (K5)

(24 / 03 / 2020, © Monfort, Dicostat2005, 2005-2020)

L'**analyse des correspondances** est une branche de l'**analyse des données** qui utilise des méthodes de **conditionnement** et de **marginalisation** d'une **loi qualitative** bivariée (cf **tableau de contingence**).

Elle se ramène à l'**analyse générale des données** par transformation d'un tableau initial.

(i) Soit $\mathcal{L}(\kappa, \lambda)$ une loi qualitative bivariée tq le **critère** qualitatif κ possède les modalités $\{1, \dots, I\}$ et le critère λ les modalités $\{1, \dots, J\}$ (cf **variable qualitative**).

On dénombre t_{ij} **unités statistiques** associées au couple de modalités i de κ et j de λ : la **loi empirique** associée est un (I, J) -tableau de contingence empirique $T \in M_{IJ}(\mathbf{N})$ à valeurs entières (**fréquences absolues**) : le terme général t_{ij} de T est ainsi une **variable de comptage**.

On appelle **analyse (factorielle) des correspondances** du tableau T l'application de l'**agd** au tableau X obtenu à partir de T par **transformation des données** (successivement) selon :

$$(1) \quad T \mapsto F = (f_{ij})_{(i,j)}, \quad \text{avec } f_{ij} = t_{ij} / t_{..} \quad (\text{fréquence relative}),$$

$$(2) \quad F \mapsto X = (x_{ij})_{(i,j)}, \quad \text{avec } x_{ij} = (f_{i.} f_{.j})^{-1/2} (f_{ij} - f_{i.} f_{.j}) \quad (\text{chi-deux élémentaire}),$$

où $t_{..} = e_i' T e_j$ (total général), la suite (finie) $(f_{1.}, \dots, f_{I.})$ représente la distribution empirique marginale associée au critère κ , et $(f_{.1}, \dots, f_{.J})$ représente la distribution empirique marginale associée à λ . Le numérateur $f_{ij} - f_{i.} f_{.j}$ de x_{ij} mesure un « **écart** » pr à l'**indépendance**, tandis que le carré du dénominateur est le produit des marginales (indépendance).

L'analyse des correspondances effectue alors la **décomposition spectrale** de $X'X$ (ou de XX').

On montre que ceci revient à diagonaliser la **matrice symétrique** S de terme général :

$$(3) \quad s_{\alpha\beta} = \sum_{i=1}^I (f_{i.})^{-1} (f_{. \alpha} f_{. \beta})^{-1/2} \cdot f_{i \alpha} f_{i \beta}.$$

(ii) L'inertie par rapport à l'origine du nuage X , ie :

$$(4) \quad \text{tr } X'X = \sum_{i=1}^I \sum_{j=1}^J x_{ij}^2,$$

n'est autre que la « **distance** » du **chi-deux** entre la distribution F des **fréquences empiriques** relatives f_{ij} (dédites de T) et la **distribution** Φ définie par le produit des marges (empiriques) $\varphi_{ij} = f_{i.} \cdot f_{.j}$, ie :

$$(5) \quad d^2(F, \Phi) = \text{tr } X' X = \sum_{i=1}^I \sum_{j=1}^J (f_{i.} f_{.j})^{-1} (f_{ij} - f_{i.} f_{.j})^2.$$

Sous l'hypothèse H_0 selon laquelle les critères du tableau T sont indépendants (ie les **variables qualitatives** κ et λ qui le définissent sont indépendantes), un **test d'indépendance** se fonde sur $d^2(F, \Phi)$ et utilise la **propriété asymptotique** :

$$(6) \quad \mathcal{L}(d^2(F, \Phi)) \xrightarrow{H_0} \min_{t(i,j) \rightarrow \infty} \mathcal{X}^2((I-1)(J-1)),$$

loi du chi-deux à $(I-1)(J-1)$ degrés de liberté.

En effet, on estime les **lois marginales** théoriques à l'aide des fréquences empiriques $e_i' F$ et $F e_j$.

(iii) Chercher la décomposition spectrale de $X' X$ équivaut à chercher celle de la matrice $Y Y'$, où $Y = (y_{ij})_{(i,j)}$ est définie par :

$$(2)' \quad y_{ij} = f_{ij} / (f_{i.} f_{.j})^{1/2}.$$

Le vecteur $u = (f_{.1}^{1/2}, \dots, f_{.J}^{1/2})'$ est le **vecteur propre** de $X' X$ relatif à la **valeur propre** nulle ; c'est aussi le vecteur propre de $Y Y'$ relatif à la valeur propre unité. C'est donc un axe factoriel « parasite » (car il correspond à une **dispersion** nulle).

La **formule de reconstitution** du tableau des fréquences F s'explique ici selon :

$$(7) \quad f_{ij} = f_{i.} f_{.j} \cdot \{1 + \sum_{\alpha=1}^{\text{rg } F} \lambda_{\alpha}^{-1/2} \cdot v_{i\alpha} w_{j\alpha}\},$$

avec :

$$(8) \quad \begin{aligned} v_{i\alpha} &= \lambda_{\alpha}^{-1/2} \sum_{j=1}^J (f_{ij} / f_{i.}) w_{j\alpha}, \\ w_{j\alpha} &= \lambda_{\alpha}^{-1/2} \sum_{i=1}^I (f_{ij} / f_{.j}) v_{i\alpha}. \end{aligned}$$

(iv) On peut étendre l'analyse au cas où les éléments t_{ij} de T sont à valeurs dans $\{0, 1\}$ ou dans $\mathbf{R}_+$. Il est parfois possible d'admettre des valeurs $t_{ij} \in \mathbf{R}$ (cases négatives), à condition que $t_{i.} > 0, \forall i$, et que $t_{.j} > 0, \forall j$, ce qui permet les transformations (1) et (2) (ou (1) et (2)').

(v) L'analyse des correspondances vérifie le **principe d'équivalence distributionnelle**, ie : si deux lignes F_{α} et F_{β} de F sont égales ($f_{\alpha i} = f_{\beta i}, \forall j$) et si on les remplace par la ligne $F_{\alpha\beta}^* = F_{\alpha} + F_{\beta}$, la distance du chi-deux entre deux colonnes quelconques f_k et f_l de F reste invariante (ie **constante**).